WASSERSTEIN AUTO-ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose the Wasserstein Auto-Encoder (WAE)—a new algorithm for building a generative model of the data distribution. WAE minimizes a penalized form of the Wasserstein distance between the model distribution and the target distribution, which leads to a different regularizer than the one used by the Variational Auto-Encoder (VAE) (Kingma & Welling, 2014). This regularizer encourages the encoded training distribution to match the prior. We compare our algorithm with several other techniques and show that it is a generalization of adversarial auto-encoders (AAE) (Makhzani et al., 2016). Our experiments show that WAE shares many of the properties of VAEs (stable training, encoder-decoder architecture, nice latent manifold structure) while generating samples of better quality.

1 INTRODUCTION

The field of representation learning was initially driven by supervised approaches, with impressive results using large labelled datasets. Unsupervised generative modeling, in contrast, used to be a domain governed by probabilistic approaches focusing on low-dimensional data. Recent years have seen a convergence of those two approaches. In the new field that formed at the intersection, variational auto-encoders (VAEs) (Kingma & Welling, 2014) form one well-established approach, theoretically elegant yet with the drawback that they tend to generate blurry samples when applied to natural images. In contrast, generative adversarial networks (GANs) (Goodfellow et al., 2014) turned out to be more impressive in terms of the visual quality of images sampled from the model, but come without an encoder, have been reported harder to train, and suffer from the mode collapse problem whereby the resulting model is unable to capture all the variability in the true data distribution. There has been a flurry of activity in assaying numerous configurations of GANs as well as combinations of VAEs and GANs. A unifying framework combining the best of GANs and VAEs in a principled way is yet to be discovered.

Following Arjovsky et al. (2017), we approach generative modeling from the optimal transport (OT) point of view. The OT cost (Villani, 2003) is a way to measure a distance between probability distributions and provides a much weaker topology than many others, including f-divergences associated with the original GAN algorithms (Nowozin et al., 2016). This is particularly important in applications, where data is usually supported on low dimensional manifolds in the input space $\mathcal{X}$. As a result, stronger notions of distances (such as f-divergences, which capture the density ratio between distributions) often max out, providing no useful gradients for training. In contrast, OT was claimed to have a nicer behaviour (Arjovsky et al., 2017; Gulrajani et al., 2017) although it requires, in its GAN-like implementation, the addition of a constraint or a regularization term into the objective.

In this work we aim at minimizing OT $W_c(P_X, P_G)$ between the true (but unknown) data distribution P_X and a *latent variable model* P_G specified by the prior distribution P_Z of latent codes $Z \in \mathcal{Z}$ and the generative model $P_G(X|Z)$ of the data points $X \in \mathcal{X}$ given Z. Our main contributions are listed below (cf. also Figure 1):

- A new family of regularized auto-encoders (Algorithms 1, 2 and Eq. 4), which we call *Wasserstein Auto-Encoders* (WAE), that minimize the optimal transport $W_c(P_X, P_G)$ for any cost function c. Similarly to VAE, the objective of WAE is composed of two terms: the c-reconstruction cost and a regularizer $\mathcal{D}_Z(P_Z, Q_Z)$ penalizing a discrepancy between two distributions in $\mathcal{Z}$: P_Z and a distribution of encoded data points, i.e. $Q_Z := \mathbb{E}_{P_X}[Q(Z|X)]$. When c is the squared cost and $\mathcal{D}_Z$ is the GAN objective, WAE coincides with adversarial auto-encoders of Makhzani et al. (2016).

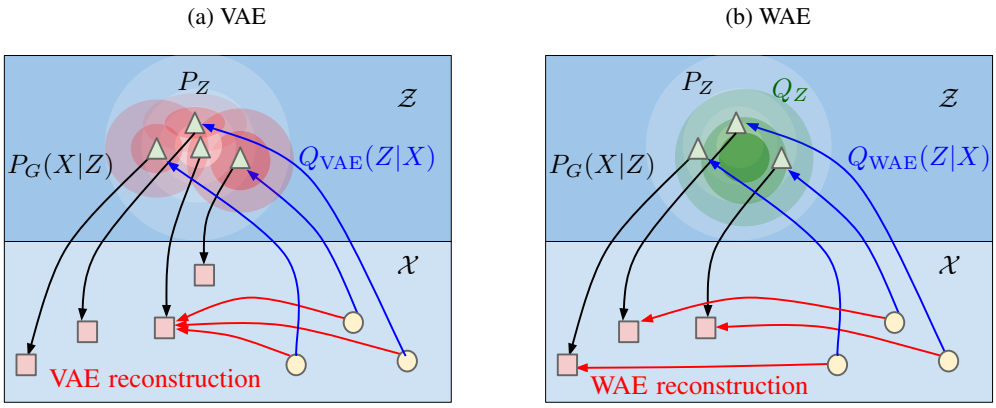

Figure 1: Both VAE and WAE minimize two terms: the reconstruction cost and the regularizer penalizing discrepancy between P_Z and distribution induced by the encoder Q. VAE forces $Q(Z|X = x)$ to match P_Z for all the different input examples x drawn from P_X. This is illustrated on picture (a), where every single red ball is forced to match P_Z depicted as the white shape. Red balls start intersecting, which leads to problems with reconstruction. In contrast, WAE forces the continuous mixture $Q_Z := \int Q(Z|X)dP_X$ to match P_Z, as depicted with the green ball in picture (b). As a result latent codes of different examples get a chance to stay far away from each other, promoting a better reconstruction.

- Empirical evaluation of WAE on MNIST and CelebA datasets with squared cost $c(x, y) = \|x - y\|_2^2$. Our experiments show that WAE keeps the good properties of VAEs (stable training, encoder-decoder architecture, and a nice latent manifold structure) while generating samples of *better quality*, approaching those of GANs.

- We propose and examine two different regularizers $\mathcal{D}_Z(P_Z, Q_Z)$. One is based on GANs and adversarial training *in the latent space* $\mathcal{Z}$. The other uses the maximum mean discrepancy, which is known to perform well when matching high-dimensional standard normal distributions P_Z (Gretton et al., 2012). Importantly, the second option leads to a fully adversary-free min-min optimization problem.

- Finally, the theoretical considerations used to derive the WAE objective might be interesting in their own right. We prove in particular (Theorem 1) that in the case of generative models, *the primal form* of $W_c(P_X, P_G)$ is equivalent to a problem involving the optimization of a probabilistic encoder $Q(Z|X)$.

The paper is structured as follows. In Section 2 we derive a novel auto-encoder formulation for OT between P_X and the latent variable model P_G. Relaxing the resulting constrained optimization problem we arrive at an objective of the Wasserstein auto-encoders. We propose two different regularizers, leading to WAE-GAN and WAE-MMD algorithms. Section 3 discusses the related work. We present the experimental results in Section 4 and conclude by pointing out some promising directions for future work.

2 Proposed method

Our new method minimizes the optimal transport cost $W_c(P_X, P_G)$ based on the novel auto-encoder formulation derived in Theorem 1. In the resulting optimization problem the decoder tries to accurately reconstruct the encoded training examples as measured by the cost function c. The encoder tries to simultaneously achieve two conflicting goals: it tries to match the encoded distribution of training examples $Q_Z := \mathbb{E}_{P_X}[Q(Z|X)]$ to the prior P_Z as measured by any specified divergence $\mathcal{D}_Z(Q_Z, P_Z)$, while making sure that the latent codes provided to the decoder are informative enough to reconstruct the encoded training examples. This is schematically depicted on Fig. 1.

2.1 PRELIMINARIES AND NOTATIONS

We use calligraphic letters (i.e. $\mathcal{X}$) for sets, capital letters (i.e. X) for random variables, and lower case letters (i.e. x) for their values. We denote probability distributions with capital letters (i.e. $P(X)$) and corresponding densities with lower case letters (i.e. $p(x)$). In this work we will consider several measures of discrepancy between probability distributions P_X and P_G. The class of *f-divergences* (Liese & Miescke, 2008) is defined by $D_f(P_X \| P_G) := \int f\big(\frac{p_X(x)}{p_G(x)}\big) p_G(x) dx$, where $f\colon (0, \infty) \to \mathcal{R}$ is any convex function satisfying $f(1) = 0$. Classical examples include the Kullback-Leibler D_{KL} and Jensen-Shannon D_{JS} divergences.

2.2 OPTIMAL TRANSPORT AND ITS DUAL FORMULATIONS

A rich class of divergences between probability distributions is induced by the *optimal transport* (OT) problem (Villani, 2003). Kantorovich's formulation of the problem is given by

$$W_c(P_X, P_G) := \inf_{\Gamma \in \mathcal{P}(X \sim P_X, Y \sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma}[c(X, Y)]\,, \tag{1}$$

where $c(x, y)\colon \mathcal{X} \times \mathcal{X} \to \mathcal{R}_+$ is any measurable *cost function* and $\mathcal{P}(X \sim P_X, Y \sim P_G)$ is a set of all joint distributions of (X, Y) with marginals P_X and P_G respectively. A particularly interesting case is when $(\mathcal{X}, d)$ is a metric space and $c(x, y) = d^p(x, y)$ for $p \geq 1$. In this case W_p, the p-th root of W_c, is called *the p-Wasserstein distance*.

When $c(x, y) = d(x, y)$ the following Kantorovich-Rubinstein duality holds[1]:

$$W_1(P_X, P_G) = \sup_{f \in \mathcal{F}_L} \mathbb{E}_{X\sim P_X}[f(X)] - \mathbb{E}_{Y\sim P_G}[f(Y)], \tag{2}$$

where $\mathcal{F}_L$ is the class of all bounded 1-Lipschitz functions on $(\mathcal{X}, d)$.

2.3 APPLICATION TO GENERATIVE MODELS: WASSERSTEIN AUTO-ENCODERS

One way to look at modern generative models like VAEs and GANs is to postulate that they are trying to minimize certain discrepancy measures between the data distribution P_X and the model P_G. Unfortunately, most of the standard divergences known in the literature, including those listed above, are hard or even impossible to compute, especially when P_X is unknown and P_G is parametrized by deep neural networks. Previous research provides several tricks to address this issue.

In case of minimizing the KL-divergence $D_{\mathrm{KL}}(P_X, P_G)$, or equivalently maximizing the marginal log-likelihood $E_{P_X}[\log p_G(X)]$, the famous *variational lower bound* provides a theoretically grounded framework successfully employed by VAEs (Kingma & Welling, 2014; Mescheder et al., 2017). More generally, if the goal is to minimize the f-divergence $D_f(P_X, P_G)$ (with one example being D_{KL}), one can resort to its dual formulation and make use of f-GANs and *the adversarial training* (Nowozin et al., 2016). Finally, OT cost $W_c(P_X, P_G)$ is yet another option, which can be, thanks to the celebrated Kantorovich-Rubinstein duality (2), expressed as an adversarial objective as implemented by the Wasserstein-GAN (Arjovsky et al., 2017). We include an extended review of all these methods in Supplementary A.

In this work we will focus on *latent variable models* P_G defined by a two-step procedure, where first a code Z is sampled from a fixed distribution P_Z on a latent space $\mathcal{Z}$ and then Z is mapped to the image $X \in \mathcal{X} = \mathcal{R}^d$ with a (possibly random) transformation. This results in a density of the form

$$p_G(x) := \int_{\mathcal{Z}} p_G(x|z) p_z(z) dz, \quad \forall x \in \mathcal{X}, \tag{3}$$

assuming all involved densities are properly defined. For simplicity we will focus on non-random decoders, i.e. generative models $P_G(X|Z)$ deterministically mapping Z to $X = G(Z)$ for a given map $G\colon \mathcal{Z} \to \mathcal{X}$. In Supplementary B we present similar results for random decoders.

It turns out that under this model, the OT cost takes a simpler form as the transportation plan factors through the map G: instead of finding a coupling Γ in (1) between two random variables living in

[1] Note that the same symbol is used for W_p and W_c, but only p is a number and thus the above W_1 refers to the 1-Wasserstein distance.

the $\mathcal{X}$ space, one distributed according to P_X and the other one according to P_G, it is sufficient to find a conditional distribution $Q(Z|X)$ such that its Z marginal $Q_Z(Z) := \mathbb{E}_{X\sim P_X}\left[Q(Z|X)\right]$ is identical to the prior distribution P_Z. This is the content of our main theorem below.

Theorem 1 *For any function $G\colon \mathcal{Z} \to \mathcal{X}$ we have*

$$\inf_{\Gamma\in\mathcal{P}(X\sim P_X, Y\sim P_G)} \mathbb{E}_{(X,Y)\sim\Gamma}\left[c\big(X,Y\big)\right] = \inf_{Q\colon Q_Z=P_Z} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\left[c\big(X,G(Z)\big)\right],$$

where Q_Z is the marginal distribution of Z when $X \sim P_X$ and $Z \sim Q(Z|X)$.

Proof The proof is reported in Supplementary B. ∎

This result allows us to optimize over random encoders $Q(Z|X)$ instead of optimizing over all couplings between X and Y. Of course, both problems are still constrained. In order to implement a numerical solution we relax the constraints on Q_Z by adding a penalty to the objective. This finally leads us to the WAE objective:

$$D_{\mathrm{WAE}}(P_X, P_G) := \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X,G(Z)\big)\big] + \lambda\cdot\mathcal{D}_Z(Q_Z, P_Z), \tag{4}$$

where $\mathcal{Q}$ is any nonparametric set of probabilistic encoders, $\mathcal{D}_Z$ is an arbitrary divergence between Q_Z and P_Z, and $\lambda > 0$ is a hyperparameter. Similarly to VAE, we propose to use deep neural networks to parametrize both encoders Q and decoders G. Note that as opposed to VAEs, the WAE formulation allows for non-random encoders deterministically mapping inputs to their latent codes.

We propose two different penalties $\mathcal{D}_Z(Q_Z, P_Z)$:

GAN-based $\mathcal{D}_Z$. The first option is to choose $\mathcal{D}_Z(Q_Z, P_Z) = D_{\mathrm{JS}}(Q_Z, P_Z)$ and use the adversarial training to estimate it. Specifically, we introduce an adversary (discriminator) in the latent space $\mathcal{Z}$ trying to separate[2] "true" points sampled from P_Z and "fake" ones sampled from Q_Z (Goodfellow et al., 2014). This results in the WAE-GAN described in Algorithm 1. Even though WAE-GAN falls back to the min-max problem, we move the adversary from the input (pixel) space $\mathcal{X}$ to the latent space $\mathcal{Z}$. On top of that, P_Z may have a nice shape with a single mode (for a Gaussian prior), in which case the task should be easier than matching an unknown, complex, and possibly multi-modal distributions as usually done in GANs. This is also a reason for our second penalty:

MMD-based $\mathcal{D}_Z$. For a positive-definite reproducing kernel $k\colon \mathcal{Z}\times\mathcal{Z}\to\mathcal{R}$ the following expression is called *the maximum mean discrepancy* (MMD):

$$\mathrm{MMD}_k(P_Z, Q_Z) = \Big\|\int_{\mathcal{Z}} k(z,\cdot)dP_Z(z) - \int_{\mathcal{Z}} k(z,\cdot)dQ_Z(z)\Big\|_{\mathcal{H}_k},$$

where $\mathcal{H}_k$ is the RKHS of real-valued functions mapping $\mathcal{Z}$ to $\mathcal{R}$. If k is *characteristic* then MMD_k defines a *metric* and can be used as a divergence measure. We propose to use $\mathcal{D}_Z(P_Z, Q_Z) = \mathrm{MMD}_k(P_Z, Q_Z)$. Fortunately, MMD has an unbiased U-statistic estimator, which can be used in conjunction with stochastic gradient descent (SGD) methods. This results in the WAE-MMD described in Algorithm 2. It is well known that the maximum mean discrepancy performs well when matching high-dimensional standard normal distributions (Gretton et al., 2012) so we expect this penalty to work especially well working with the Gaussian prior P_Z.

3 Related work

Literature on auto-encoders Classical unregularized auto-encoders minimize only the reconstruction cost. This results in different training points being encoded into non-overlapping zones chaotically scattered all across the $\mathcal{Z}$ space with "holes" in between where the decoder mapping $P_G(X|Z)$ has never been trained. Overall, the encoder $Q(Z|X)$ trained in this way does not provide a useful representation and sampling from the latent space $\mathcal{Z}$ becomes hard (Bengio et al., 2013).

Variational auto-encoders (Kingma & Welling, 2014) minimize a variational bound on the KL-divergence $D_{\mathrm{KL}}(P_X, P_G)$ which is composed of the reconstruction cost plus

[2]We noticed that the famous "log trick" (also called "non saturating loss") proposed by Goodfellow et al. (2014) leads to better results.

$\mathbb{E}_{P_X}\left[D_{\mathrm{KL}}(Q(Z|X), P_Z)\right]$ which captures how distinct the image by the encoder of *each* training example is from the prior P_Z, which is not guaranteeing that the overall encoded distribution $\mathbb{E}_{P_X}\left[Q(Z|X)\right]$ matches P_Z like WAE does. Also, VAEs require non-degenerate Gaussian encoders and random decoders for which $\log p_G(x|z)$ can be computed and differentiated with respect to the parameters. Later Mescheder et al. (2017) proposed a way to use VAE with non-Gaussian encoders. WAE minimizes OT $W_c(P_X, P_G)$ and allows both probabilistic and deterministic encoder-decoder pairs of any kind.

ALGORITHM 1 Wasserstein Auto-Encoder with GAN-based penalty (WAE-GAN).

Require: Regularization coefficient $\lambda > 0$.
Initialize the parameters of the encoder Q_ϕ, decoder G_θ, and latent discriminator D_γ.
while (ϕ, θ) not converged **do**
Sample $\{x_1, \dots, x_n\}$ from the training set
Sample $\{z_1, \dots, z_n\}$ from the prior P_Z
Sample $\tilde{z}_i$ from $Q_\phi(Z|x_i)$ for $i = 1, \dots, n$
Update D_γ by ascending:

$$\frac{\lambda}{n} \sum_{i=1}^{n} \log D_\gamma(z_i) + \log\big(1 - D_\gamma(\tilde{z}_i)\big)$$

Update Q_ϕ and G_θ by descending:

$$\frac{1}{n} \sum_{i=1}^{n} c\big(x_i, G_\theta(\tilde{z}_i)\big) - \lambda \cdot \log D_\gamma(\tilde{z}_i)$$

end while

ALGORITHM 2 Wasserstein Auto-Encoder with MMD-based penalty (WAE-MMD).

Require: Regularization coefficient $\lambda > 0$, characteristic positive-definite kernel k.
Initialize the parameters of the encoder Q_ϕ, decoder G_θ, and latent discriminator D_γ.
while (ϕ, θ) not converged **do**
Sample $\{x_1, \dots, x_n\}$ from the training set
Sample $\{z_1, \dots, z_n\}$ from the prior P_Z
Sample $\tilde{z}_i$ from $Q_\phi(Z|x_i)$ for $i = 1, \dots, n$
Update Q_ϕ and G_θ by descending:

$$\frac{1}{n} \sum_{i=1}^{n} c\big(x_i, G_\theta(\tilde{z}_i)\big) + \frac{\lambda}{n(n-1)} \sum_{\ell \neq j} k(z_\ell, z_j) + k(\tilde{z}_\ell, \tilde{z}_j) - 2k(z_\ell, \tilde{z}_j)$$

end while

When used with $c(x, y) = \|x - y\|_2^2$ WAE-GAN is equivalent to adversarial auto-encoders (AAE) proposed by Makhzani et al. (2016). Our theory thus suggests that AAEs minimize the 2-Wasserstein distance between P_X and P_G. This provides the first theoretical justification for AAEs known to the authors. WAE generalizes AAE in two ways: first, it can use any cost function c in the input space $\mathcal{X}$; second, it can use any discrepancy measure $\mathcal{D}_Z$ in the latent space $\mathcal{Z}$ (for instance MMD), not necessarily the adversarial one of WAE-GAN.

Literature on OT Genevay et al. (2016) address computing the OT cost in large scale using SGD and sampling. They approach this task either through the dual formulation, or via a regularized version of the primal. They do not discuss any implications for generative modeling. Our approach is based on the primal form of OT, we arrive at regularizers which are very different, and our main focus is on generative modeling.

The WGAN (Arjovsky et al., 2017) minimizes the 1-Wasserstein distance $W_1(P_X, P_G)$ for generative modeling. The authors approach this task from the dual form. Their algorithm comes without an encoder and can not be readily applied to any other cost W_c, because the neat form of the Kantorovich-Rubinstein duality (2) holds only for W_1. WAE approaches the same problem from the primal form, can be applied for any cost function c, and comes naturally with an encoder.

In order to compute the values (1) or (2) of OT we need to handle non-trivial constraints, either on the coupling distribution Γ or on the function f being considered. Various approaches have been proposed in the literature to circumvent this difficulty. For W_1 Arjovsky et al. (2017) tried to implement the constraint in the dual formulation (2) by clipping the weights of the neural network f. Later Gulrajani et al. (2017) proposed to relax the same constraint by penalizing the objective of (2) with a term $\lambda \cdot \mathbb{E}\left(\|\nabla f(X)\| - 1\right)^2$ which should not be greater than 1 if $f \in \mathcal{F}_L$. In a more general OT setting of W_c Cuturi (2013) proposed to penalize the objective of (1) with the KL-divergence $\lambda \cdot D_{\mathrm{KL}}(\Gamma, P \otimes Q)$ between the coupling distribution and the product of marginals. Genevay et al. (2016) showed that this entropic regularization drops the constraints on functions in the dual formulation as opposed to (2). Finally, in the context of *unbalanced optimal transport* it

has been proposed to relax the constraint in (1) by regularizing the objective with $\lambda \cdot \big(D_f(\Gamma_X, P) + D_f(\Gamma_Y, Q)\big)$ (Chizat et al., 2015; Liero et al., 2015), where Γ_X and Γ_Y are marginals of Γ. In this paper we propose to relax OT in a way similar to the unbalanced optimal transport, i.e. by adding additional divergences to the objective. However, we show that in the particular context of generative modeling, only one extra divergence is necessary.

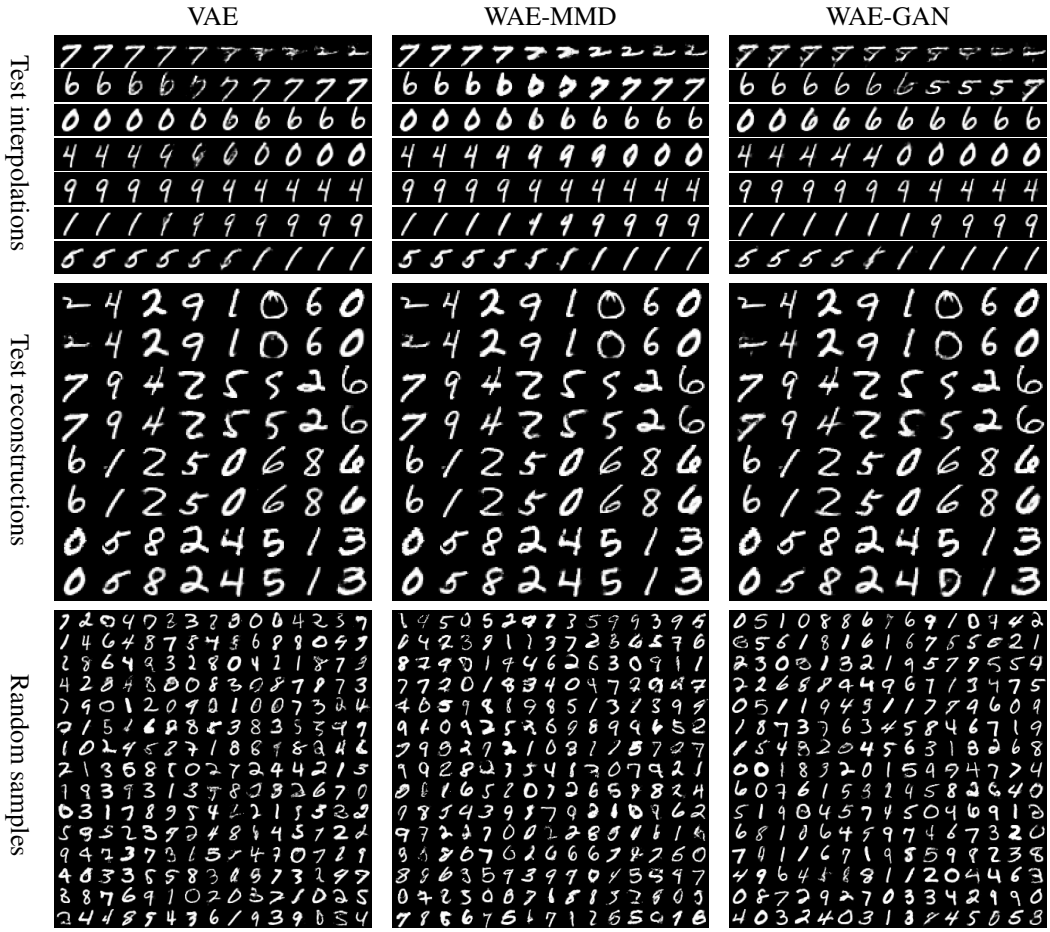

Figure 2: VAE (left column), WAE-MMD (middle column), and WAE-GAN (right column) trained on MNIST dataset. In "test reconstructions" odd rows correspond to the real test points.

Literature on GANs Many of the GAN variations (including f-GAN and WGAN) come without an encoder. Often it may be desirable to reconstruct the latent codes and use the learned manifold, in which cases these models are not applicable.

There have been many other approaches trying to blend the adversarial training of GANs with auto-encoder architectures (Zhao et al., 2017; Dumoulin et al., 2017; Ulyanov et al., 2017; Berthelot et al., 2017). The approach proposed by Ulyanov et al. (2017) is perhaps the most relevant to our work. The authors use the discrepancy between Q_Z and the distribution $\mathbb{E}_{Z' \sim P_Z}[Q\big(Z|G(Z')\big)]$ of auto-encoded noise vectors as the objective for the max-min game between the encoder and decoder respectively. While the authors showed that the saddle points correspond to $P_X = P_G$, they admit that encoders and decoders trained in this way have no incentive to be reciprocal. As a workaround they propose to include an additional reconstruction term to the objective. WAE does not necessarily lead to a min-max game, uses a different penalty, and has a clear theoretical foundation.

Several works used reproducing kernels in context of GANs. Li et al. (2015); Dziugaite et al. (2015) use MMD with a fixed kernel k to match P_X and P_G directly in the input space $\mathcal{X}$. These methods have been criticised to require larger mini-batches during training: estimating $\mathrm{MMD}_k(P_X, P_G)$ requires number of samples roughly proportional to the dimensionality of the input space $\mathcal{X}$ (Reddi

et al., 2015) which is typically larger than 10^3. Li et al. (2017) take a similar approach but further train k adversarially so as to arrive at a meaningful loss function. WAE-MMD uses MMD to match Q_Z to the prior P_Z in the latent space $\mathcal{Z}$. Typically $\mathcal{Z}$ has no more than 100 dimensions and P_Z is Gaussian, which allows us to use regular mini-batch sizes to accurately estimate MMD.

Figure 3: VAE (left column), WAE-MMD (middle column), and WAE-GAN (right column) trained on CelebA dataset. In "test reconstructions" odd rows correspond to the real test points.

4 Experiments

In this section we empirically evaluate the proposed WAE model. We would like to test if WAE can simultaneously achieve (i) accurate reconstructions of data points, (ii) reasonable geometry of the latent manifold, and (iii) random samples of good (visual) quality. Importantly, the model should generalize well: requirements (i) and (ii) should be met on both training and test data. We trained WAE-GAN and WAE-MMD (Algorithms 1 and 2) on two real-world datasets: MNIST (LeCun et al., 1998) consisting of 70k images and CelebA (Liu et al., 2015) containing roughly 203k images.

Experimental setup In all reported experiments we used Euclidian latent spaces $\mathcal{Z} = \mathcal{R}^{d_z}$ for various d_z depending on the complexity of the dataset, isotropic Gaussian prior distributions $P_Z(Z) = \mathcal{N}(Z; \mathbf{0}, \sigma_z^2 \cdot \boldsymbol{I}_d)$ over $\mathcal{Z}$, and a squared cost function $c(x, y) = \|x - y\|_2^2$ for data points $x, y \in \mathcal{X} = \mathcal{R}^{d_x}$. We used *deterministic* encoder-decoder pairs, Adam (Kingma & Lei, 2014) with $\beta_1 = 0.5, \beta_2 = 0.999$, and convolutional deep neural network architectures for encoder mapping $Q_\phi \colon \mathcal{X} \to \mathcal{Z}$ and decoder mapping $G_\theta \colon \mathcal{Z} \to \mathcal{X}$ similar to the DCGAN ones reported by Radford et al. (2016) with batch normalization (Ioffe & Szegedy, 2015). We tried various values of λ and noticed that $\lambda = 10$ seems to work good across all datasets we considered. All reported experiments use this value.

Since we are using deterministic encoders, choosing d_z larger than intrinsic dimensionality of the dataset would force the encoded distribution Q_Z to live on a manifold in $\mathcal{Z}$. This would make matching Q_Z to P_Z impossible if P_Z is Gaussian and may lead to numerical instabilities. We use $d_z = 8$ for MNIST and $d_z = 64$ for CelebA which seems to work reasonably well.

We also report results of VAEs. VAEs used the same latent spaces as discussed above and standard Gaussian priors $P_Z = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. We used Gaussian encoders $Q(Z|X) = \mathcal{N}\big(Z; Q_\phi(X), \Sigma(X)\big)$ with mean Q_ϕ and diagonal covariance Σ and Gaussian decoders $P_G(X|Z) = \mathcal{N}\big(Z; G_\theta(X), \sigma_G^2 \cdot \boldsymbol{I}_d\big)$ with mean $G_\theta(Z)$. Functions Q_ϕ, Σ, and G_θ were parametrized by deep nets of the same architectures as used in WAE.

WAE-GAN and WAE-MMD specifics In WAE-GAN we used discriminator D composed of several fully connected layers with ReLu. We tried WAE-MMD with the RBF kernel but observed that it fails to penalize the outliers of Q_Z because of the quick tail decay. If the codes $\tilde{z} = Q_\phi(x)$ for some of the training points $x \in \mathcal{X}$ end up far away from the support of P_Z (which may happen in the early stages of training) the corresponding terms in the U-statistic $k(z, \tilde{z}) = e^{-\|\tilde{z}-z\|_2^2/\sigma_k^2}$ will quickly approach zero and provide no gradient for those outliers. This could be avoided by choosing the kernel bandwidth σ_k^2 in a data-dependent manner, however in this case per-minibatch U-statistic would not provide an unbiased estimate for the gradient. Instead, we used the *inverse multiquadratics* kernel $k(x, y) = C/(C + \|x - y\|_2^2)$ which is also characteristic and has much heavier tails. In all experiments we used $C = 2d_z\sigma_z^2$, which is the expected squared distance between two multivariate Gaussian vectors drawn from P_Z. This significantly improved the performance compared to the RBF kernel (even the one with $\sigma_k^2 = 2d_z\sigma_z^2$).

Trained models are presented in Figures 2 and 3. Further details are presented in Supplementary E.

Random samples are generated by sampling P_Z and decoding the resulting noise vectors z into $G_\theta(z)$. As expected, in our experiments we observed that for both WAE-GAN and WAE-MMD the quality of samples strongly depends on how accurately Q_Z matches P_Z. To see this, notice that while training the decoder function G_θ is presented only with encoded versions $Q_\phi(X)$ of the data points $X \sim P_X$. In other words, the decoder is trained on samples drown from Q_Z and thus there is no reason to expect good results when feeding it with samples from P_Z. We noticed that even slightest differences between Q_Z and P_Z may affect the quality of samples.

Test reconstructions and interpolations. We take random points x from the held out test set and report their auto-encoded versions $G_\theta(Q_\phi(x))$. Next, pairs (x, y) of different data points are sampled randomly from the held out test set and encoded: $z_x = Q_\phi(x)$, $z_y = Q_\phi(y)$. We *linearly* interpolate between z_x and z_y with equally-sized steps in the latent space and report the results of decoding.

5 Conclusion

Using the optimal transport cost, we have derived the Wasserstein auto-encoders—a new family of algorithms for building generative models. We discussed their relations to other probabilistic modeling techniques. We conducted experiments using two particular implementations of the proposed method showing that in comparison to VAE, the images sampled from the trained WAE models are of better quality, without compromising the stability of training and the quality of reconstruction. Future work will include further exploration of the criteria for matching the encoded distribution Q_Z to the prior distribution P_Z, testing the possibility of adversarially training the cost function c in the input space $\mathcal{X}$, and a theoretical analysis of the dual formulations for WAE-GAN and WAE-MMD.

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN, 2017.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence*, 35, 2013.

D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks, 2017.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: geometry and kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. In *ICLR*, 2017.

G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

A. Genevay, M. Cuturi, G. Peyré, and F. R. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pp. 3432–3440, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Domoulin, and A. Courville. Improved training of wasserstein GANs, 2017.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

D. P. Kingma and J. Lei. Adam: A method for stochastic optimization, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86(11), pp. 2278–2324, 1998.

C. L. Li, W. C. Chang, Y. Cheng, Y. Yang, and B. Poczos. Mmd gan: Towards deeper understanding of moment matching network, 2017.

Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *ICML*, 2015.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger-kantorovich distance between positive measures. *arXiv preprint arXiv:1508.07941*, 2015.

F. Liese and K.-J. Miescke. *Statistical Decision Theory*. Springer, 2008.

J. Lin. Divergence measures based on the shannon entropy. *Information Theory*, 37, 1991.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow. Adversarial autoencoders. In *ICLR*, 2016.

L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks, 2017.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016.

B. Poole, A. Alemi, J. Sohl-Dickstein, and A. Angelova. Improved generator objectives for GANs, 2016.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

R. Reddi, A. Ramdas, A. Singh, B. Poczos, and L. Wasserman. On the high-dimensional power of a linear-time two sample test under mean-shift alternatives. In *AISTATS*, 2015.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, NY, 2008.

D. Ulyanov, A. Vedaldi, and V. Lempitsky. It takes (only) two: Adversarial generator-encoder networks, 2017.

C. Villani. *Topics in Optimal Transportation*. AMS Graduate Studies in Mathematics, 2003.

J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. In *ICLR*, 2017.

A IMPLICIT GENERATIVE MODELS: A SHORT TOUR OF GANS AND VAES

Even though GANs and VAEs are quite different—both in terms of the conceptual frameworks and empirical performance—they share important features: (a) both can be trained by sampling from the model P_G without knowing an analytical form of its density and (b) both can be scaled up with SGD. As a result, it becomes possible to use highly flexible *implicit* models P_G defined by a two-step procedure, where first a code Z is sampled from a fixed distribution P_Z on a latent space $\mathcal{Z}$ and then Z is mapped to the image $G(Z) \in \mathcal{X} = \mathcal{R}^d$ with a (possibly random) transformation $G\colon \mathcal{Z} \to \mathcal{X}$. This results in *latent variable models* P_G of the form (3).

These models are indeed easy to sample and, provided G can be differentiated analytically with respect to its parameters, P_G can be trained with SGD. The field is growing rapidly and numerous variations of VAEs and GANs are available in the literature. Next we introduce and compare several of them.

The original **generative adversarial network** (GAN) Goodfellow et al. (2014) approach minimizes

$$D_{\mathrm{GAN}}(P_X, P_G) = \sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim P_X}[\log T(X)] + \mathbb{E}_{Z \sim P_Z}\big[\log\big(1 - T(G(Z))\big)\big] \tag{5}$$

with respect to a deterministic *decoder* $G\colon \mathcal{Z} \to \mathcal{X}$, where $\mathcal{T}$ is any non-parametric class of choice. It is known that $D_{\mathrm{GAN}}(P_X, P_G) \leq 2 \cdot D_{\mathrm{JS}}(P_X, P_G) - \log(4)$ and the inequality turns into identity in the *nonparametric limit*, that is when the class $\mathcal{T}$ becomes rich enough to represent *all* functions mapping $\mathcal{X}$ to $(0, 1)$. Hence, GANs are *minimizing a lower bound* on the JS-divergence. However, GANs are not only linked to the JS-divergence: the f-GAN approach Nowozin et al. (2016) showed that a slight modification $D_{\mathrm{f,GAN}}$ of the objective (5) allows to lower bound any desired f-divergence in a similar way. In practice, both decoder G and *discriminator* T are trained in alternating SGD steps. Stopping criteria as well as adequate evaluation of the trained GAN models remain open questions.

Recently, the authors of Arjovsky et al. (2017) argued that the 1-Wasserstein distance W_1, which is known to induce a much weaker topology than D_{JS}, may be better suited for generative modeling. When P_X and P_G are supported on largely disjoint low-dimensional manifolds (which may be the case in applications), D_{KL}, D_{JS}, and other strong distances between P_X and P_G max out and no longer provide useful gradients for P_G. This "vanishing gradient" problem necessitates complicated scheduling between the G/T updates. In contrast, W_1 is still sensible in these cases and provides stable gradients. The **Wasserstein GAN** (WGAN) minimizes

$$D_{\mathrm{WGAN}}(P_X, P_G) = \sup_{T \in \mathcal{W}} \mathbb{E}_{X \sim P_X}[T(X)] - \mathbb{E}_{Z \sim P_Z}\big[T(G(Z))\big],$$

where $\mathcal{W}$ is any subset of 1-Lipschitz functions on $\mathcal{X}$. It follows from (2) that $D_{\mathrm{WGAN}}(P_X, P_G) \leq W_1(P_X, P_G)$ and thus WGAN is *minimizing a lower bound* on the 1-Wasserstein distance.

Variational auto-encoders (VAE) Kingma & Welling (2014) utilize models P_G of the form (3) and minimize

$$D_{\mathrm{VAE}}(P_X, P_G) = \inf_{Q(Z|X) \in \mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_Z\big) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] \tag{6}$$

with respect to a random *decoder* mapping $P_G(X|Z)$. The conditional distribution $P_G(X|Z)$ is often parametrized by a deep net G and can have any form as long as its density $p_G(x|z)$ can be computed and differentiated with respect to the parameters of G. A typical choice is to use Gaussians $P_G(X|Z) = \mathcal{N}(X; G(Z), \sigma^2 \cdot I)$. If $\mathcal{Q}$ is the set of *all* conditional probability distributions $Q(Z|X)$, the objective of VAE coincides with the negative marginal log-likelihood $D_{\mathrm{VAE}}(P_X, P_G) = -\mathbb{E}_{P_X}[\log P_G(X)]$. However, in order to make the D_{KL} term of (6) tractable in closed form, the original implementation of VAE uses a standard normal P_Z and restricts $\mathcal{Q}$ to a class of Gaussian distributions $Q(Z|X) = \mathcal{N}\big(Z; \mu(X), \Sigma(X)\big)$ with mean μ and diagonal covariance Σ parametrized by deep nets. As a consequence, VAE is *minimizing an upper bound* on the negative log-likelihood or, equivalently, on the KL-divergence $D_{\mathrm{KL}}(P_X, P_G)$. Further details can be found in Section D.

One possible way to reduce the gap between the true negative log-likelihood and the upper bound provided by D_{VAE} is to enlarge the class $\mathcal{Q}$. **Adversarial variational Bayes** (AVB) Mescheder et al. (2017) follows this argument by employing the idea of GANs. Given any point $x \in \mathcal{X}$,

a noise $\epsilon \sim \mathcal{N}(0,1)$, and any fixed transformation $e\colon \mathcal{X} \times \mathcal{R} \to \mathcal{Z}$, a random variable $e(x,\epsilon)$ implicitly defines one particular conditional distribution $Q_e(Z|X=x)$. AVB allows $\mathcal{Q}$ to contain all such distributions for different choices of e, replaces the intractable term $D_{\mathrm{KL}}\big(Q_e(Z|X), P_Z\big)$ in (6) by the adversarial approximation $D_{\mathrm{f,GAN}}$ corresponding to the KL-divergence, and proposes to minimize[3]

$$D_{\mathrm{AVB}}(P_X, P_G) = \inf_{Q_e(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{f,GAN}}\big(Q_e(Z|X), P_Z\big) - \mathbb{E}_{Q_e(Z|X)}[\log p_G(X|Z)]\right]. \quad (7)$$

The D_{KL} term in (6) may be viewed as a regularizer. Indeed, VAE reduces to the classical unregularized auto-encoder if this term is dropped, minimizing the reconstruction cost of the encoder-decoder pair $Q(Z|X), P_G(X|Z)$. This often results in different training points being encoded into non-overlapping zones chaotically scattered all across the $\mathcal{Z}$ space with "holes" in between where the decoder mapping $P_G(X|Z)$ has never been trained. Overall, the encoder $Q(Z|X)$ trained in this way does not provide a useful representation and sampling from the latent space $\mathcal{Z}$ becomes hard Bengio et al. (2013).

Adversarial auto-encoders (AAE) Makhzani et al. (2016) replace the D_{KL} term in (6) with another regularizer:

$$D_{\mathrm{AAE}}(P_X, P_G) = \inf_{Q(Z|X)\in\mathcal{Q}} D_{\mathrm{GAN}}(Q_Z, P_Z) - \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)], \quad (8)$$

where Q_Z is the marginal distribution of Z when first X is sampled from P_X and then Z is sampled from $Q(Z|X)$, also known as the *aggregated posterior* Makhzani et al. (2016). Similarly to AVB, there is no clear link to log-likelihood, as $D_{\mathrm{AAE}} \leq D_{\mathrm{AVB}}$ (see Appendix D). The authors of Makhzani et al. (2016) argue that matching Q_Z to P_Z in this way ensures that there are no "holes" left in the latent space $\mathcal{Z}$ and $P_G(X|Z)$ generates reasonable samples whenever $Z \sim P_Z$. They also report an equally good performance of different types of conditional distributions $Q(Z|X)$, including Gaussians as used in VAEs, implicit models Q_e as used in AVB, and *deterministic* encoder mappings, i.e. $Q(Z|X) = \delta_{\mu(X)}$ with $\mu\colon \mathcal{X} \to \mathcal{Z}$.

B Proof of Theorem 1 and further details

We will consider certain sets of joint probability distributions of three random variables $(X, Y, Z) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Z}$. The reader may wish to think of X as true images, Y as images sampled from the model, and Z as latent codes. We denote by $P_{G,Z}(Y,Z)$ a joint distribution of a variable pair (Y,Z), where Z is first sampled from P_Z and next Y from $P_G(Y|Z)$. Note that P_G defined in (3) and used throughout this work is the marginal distribution of Y when $(Y,Z) \sim P_{G,Z}$.

In the optimal transport problem (1), we consider joint distributions $\Gamma(X,Y)$ which are called *couplings* between values of X and Y. Because of the marginal constraint, we can write $\Gamma(X,Y) = \Gamma(Y|X)P_X(X)$ and we can consider $\Gamma(Y|X)$ as a non-deterministic mapping from X to Y. Theorem 1. shows how to *factor* this mapping through $\mathcal{Z}$, i.e., decompose it into an encoding distribution $Q(Z|X)$ and the generating distribution $P_G(Y|Z)$.

As in Section 2.2, $\mathcal{P}(X \sim P_X, Y \sim P_G)$ denotes the set of all joint distributions of (X,Y) with marginals P_X, P_G, and likewise for $\mathcal{P}(X \sim P_X, Z \sim P_Z)$. The set of all joint distributions of (X,Y,Z) such that $X \sim P_X$, $(Y,Z) \sim P_{G,Z}$, and $(Y \perp\!\!\!\perp X)|Z$ will be denoted by $\mathcal{P}_{X,Y,Z}$. Finally, we denote by $\mathcal{P}_{X,Y}$ and $\mathcal{P}_{X,Z}$ the sets of marginals on (X,Y) and (X,Z) (respectively) induced by distributions in $\mathcal{P}_{X,Y,Z}$. Note that $\mathcal{P}(P_X, P_G)$, $\mathcal{P}_{X,Y,Z}$, and $\mathcal{P}_{X,Y}$ depend on the choice of conditional distributions $P_G(Y|Z)$, while $\mathcal{P}_{X,Z}$ does not. In fact, it is easy to check that $\mathcal{P}_{X,Z} = \mathcal{P}(X \sim P_X, Z \sim P_Z)$. From the definitions it is clear that $\mathcal{P}_{X,Y} \subseteq \mathcal{P}(P_X, P_G)$ and we immediately get the following upper bound:

$$W_c(P_X, P_G) \leq W_c^{\dagger}(P_X, P_G) := \inf_{P\in\mathcal{P}_{X,Y}} \mathbb{E}_{(X,Y)\sim P}\left[c(X,Y)\right]. \quad (9)$$

If $P_G(Y|Z)$ are Dirac measures (i.e., $Y = G(Z)$), it turns out that $\mathcal{P}_{X,Y} = \mathcal{P}(P_X, P_G)$:

[3]The authors of AVB Mescheder et al. (2017) note that using f-GAN as described above actually results in "unstable training". Instead, following the approach of Poole et al. (2016), they use a trained discriminator T^* resulting from the D_{GAN} objective (5) to approximate the ratio of densities and then directly estimate the KL divergence $\int f\big(p(x)/q(x)\big)q(x)dx$.

Lemma 1 *$\mathcal{P}_{X,Y} \subseteq \mathcal{P}(P_X, P_G)$ with identity if*[4] *$P_G(Y|Z=z)$ are Dirac for all $z \in \mathcal{Z}$.*

Proof The first assertion is obvious. To prove the identity, note that when Y is a deterministic function of Z, for any A in the sigma-algebra induced by Y we have $\mathbb{E}\left[\mathbf{1}_{[Y\in A]}|X,Z\right] = \mathbb{E}\left[\mathbf{1}_{[Y\in A]}|Z\right]$. This implies $(Y \perp\!\!\!\perp X)|Z$ and concludes the proof. ∎

We are now in place to prove Theorem 1. Lemma 1 obviously leads to

$$W_c(P_X, P_G) = W_c^{\dagger}(P_X, P_G).$$

The tower rule of expectation, and the conditional independence property of $\mathcal{P}_{X,Y,Z}$ implies

$$\begin{aligned}
W_c^{\dagger}(P_X, P_G) &= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{(X,Y,Z)\sim P}\left[c(X,Y)\right] \\
&= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{P_Z}\mathbb{E}_{X\sim P(X|Z)}\mathbb{E}_{Y\sim P(Y|Z)}[c(X,Y)] \\
&= \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{P_Z}\mathbb{E}_{X\sim P(X|Z)}\left[c\big(X, G(Z)\big)\right] \\
&= \inf_{P\in\mathcal{P}_{X,Z}} \mathbb{E}_{(X,Z)\sim P}\left[c\big(X, G(Z)\big)\right].
\end{aligned}$$

It remains to notice that $\mathcal{P}_{X,Z} = \mathcal{P}(X \sim P_X, Z \sim P_Z)$ as stated earlier.

B.1 Random decoders $P_G(Y|Z)$

If the decoders are non-deterministic, Lemma 1 provides only the inclusion of sets $\mathcal{P}_{X,Y} \subseteq \mathcal{P}(P_X, P_G)$ and we get the following upper bound on the OT:

Corollary 1 *Let $\mathcal{X} = \mathcal{R}^d$ and assume the conditional distributions $P_G(Y|Z=z)$ have mean values $G(z) \in \mathcal{R}^d$ and marginal variances $\sigma_1^2, \dots, \sigma_d^2 \geq 0$ for all $z \in \mathcal{Z}$, where $G\colon \mathcal{Z} \to \mathcal{X}$. Take $c(x,y) = \|x-y\|_2^2$. Then*

$$W_c(P_X, P_G) \leq W_c^{\dagger}(P_X, P_G) = \sum_{i=1}^{d} \sigma_i^2 + \inf_{P\in\mathcal{P}(X\sim P_X, Z\sim P_Z)} \mathbb{E}_{(X,Z)\sim P}\left[\|X - G(Z)\|^2\right]. \quad (10)$$

Proof First inequality follows from (9). For the identity we proceed similarly to the proof of Theorem 1 and write

$$W_c^{\dagger}(P_X, P_G) = \inf_{P\in\mathcal{P}_{X,Y,Z}} \mathbb{E}_{P_Z}\mathbb{E}_{X\sim P(X|Z)}\mathbb{E}_{Y\sim P(Y|Z)}\left[\|X-Y\|^2\right]. \quad (11)$$

Note that

$$\begin{aligned}
&\mathbb{E}_{Y\sim P(Y|Z)}\left[\|X-Y\|^2\right] = \mathbb{E}_{Y\sim P(Y|Z)}\left[\|X - G(Z) + G(Z) - Y\|^2\right] \\
&= \|X - G(Z)\|^2 + \mathbb{E}_{Y\sim P(Y|Z)}\left[\langle X - G(Z), G(Z) - Y\rangle\right] + \mathbb{E}_{Y\sim P(Y|Z)}\|G(Z) - Y\|^2 \\
&= \|X - G(Z)\|^2 + \sum_{i=1}^{d}\sigma_i^2.
\end{aligned}$$

Together with (11) and the fact that $\mathcal{P}_{X,Z} = \mathcal{P}(X \sim P_X, Z \sim P_Z)$ this concludes the proof. ∎

C Relations of WAE to AAE, VAEs, and GANs

In Section C.1, we compare minimizing the optimal transport cost W_c, the upper bound $W_c^{\dagger}$, and its relaxed version D_{WAE} to VAE, AVB, and AAE in the special case when $c(x,y) = \|x-y\|^2$ and $P_G(Y|Z) = \mathcal{N}(Y; G(Z), \sigma^2 \cdot I)$. We show that in this case the solutions of VAE and AVB both

[4] We conjecture that this is also a necessary condition. The necessity is not used in the paper.

depend on σ^2, while the minimizer $G^\dagger$ of $W_c^\dagger(P_X, P_G)$ does not depend on σ^2, and is the same as the minimizer of $W_c(P_X, P_G)$ for $\sigma^2 = 0$. We also briefly discuss the role of these conclusions in explaining the well-known blurriness of VAE outputs. Section C.2 shows that when $c(x, y) = \|x - y\|$, WAE and WGAN approach primal and dual forms respectively of the *same optimization problem*. Finally, we discuss a difference in behaviour between the two algorithms caused by this duality.

We refer the reader to Supplementary A for the detailed overview of all these methods.

C.1 THE 2-WASSERSTEIN DISTANCE: RELATION TO VAE, AVB, AND AAE

Consider the squared Euclidean cost function $c(x, y) = \|x - y\|^2$, for which W_c is the squared 2-Wasserstein distance W_2^2. The goal of this section is to compare the minimization of $W_2(P_X, P_G)$ to other generative modeling approaches. Let us focus our attention on generative distributions $P_G(Y|Z)$ typically used in VAE, AVB, and AAE, i.e., Gaussians $P_G(Y|Z) = \mathcal{N}\big(Y; G(Z), \sigma^2 \cdot I_d\big)$. In order to verify the differentiability of $\log p_G(x|z)$ all three methods require $\sigma^2 > 0$ and have problems handling the case of deterministic decoders ($\sigma^2 = 0$). To emphasize the role of the variance σ^2 we will denote the resulting latent variable model P_G^σ.

Relation to VAE and AVB The analysis of Supplementary B shows that the value of $W_2(P_X, P_G^\sigma)$ is upper bounded by $W_c^\dagger(P_X, P_G^\sigma)$ of the form (10) and the two coincide when $\sigma^2 = 0$. Next we summarize properties of solutions G minimizing these two values W_2 and $W_c^\dagger$:

Proposition 1 *Let $\mathcal{X} = \mathcal{R}^d$ and assume $c(x, y) = \|x - y\|^2$, $P_G(Y|Z) = \mathcal{N}\big(Y; G(Z), \sigma^2 \cdot I\big)$ with any function $G\colon \mathcal{X} \to \mathcal{R}$. If $\sigma^2 > 0$ then the functions G_σ^* and $G^\dagger$ minimizing $W_c(P_X, P_G^\sigma)$ and $W_c^\dagger(P_X, P_G^\sigma)$ respectively are different: G_σ^* depends on σ^2, while $G^\dagger$ does not. The function $G^\dagger$ is also a minimizer of $W_c(P_X, P_G^0)$.*

In order to prove this proposition we will need the following simple result, which is basically saying that the variance of a sum of two independent random variables is a sum of the variances:

Lemma 2 *Under conditions of Proposition 1, assume $Y \sim P_G^\sigma$. Then*

$$\mathrm{Var}[Y] = \sigma^2 + \mathrm{Var}_{Z \sim P_Z}[G(Z)].$$

Proof First of all, using (3) we have

$$\mathbb{E}[Y] := \int_{\mathcal{R}} y \int_{\mathcal{Z}} p_G(y|z) p_Z(z) dz dy = \int_{\mathcal{Z}} \left(\int_{\mathcal{R}} y \, p_G(y|z) dy \right) p_Z(z) dz = \mathbb{E}_{Z \sim P_Z}[G(Z)].$$

Then

$$\begin{aligned}
\mathrm{Var}[Y] &:= \int_{\mathcal{R}} (y - \mathbb{E}[G(Z)])^2 \int_{\mathcal{Z}} p_G(y|z) p_Z(z) dz dy \\
&= \int_{\mathcal{R}} (y - G(z))^2 \int_{\mathcal{Z}} p_G(y|z) p_Z(z) dz dy + \int_{\mathcal{R}} (G(z) - \mathbb{E}[G(Z)])^2 \int_{\mathcal{Z}} p_G(y|z) p_Z(z) dz dy \\
&= \sigma^2 + \mathrm{Var}_{Z \sim P_Z}[G(Z)].
\end{aligned}$$

∎

Next we turn to the proof of Proposition 1:

Proof Corollary 1 shows that $G^\dagger$ does not depend on the variance σ^2. When $\sigma^2 = 0$ the distribution $P_G(Y|Z)$ turns into Dirac. In this case we combine Theorem 1 and Corollary 1 to conclude that $G^\dagger$ also minimizes $W_c(P_X, P_G^0)$. Next we prove that when $\sigma^2 > 0$ the function G_σ^* minimizing $W_c(P_X, P_G^\sigma)$ depends on σ^2.

The proof is based on the following example: $\mathcal{X} = \mathcal{Z} = \mathcal{R}$, $P_X = \mathcal{N}(0, 1)$, $P_Z = \mathcal{N}(0, 1)$, and $0 < \sigma^2 < 1$. Note that by setting $G(z) = c \cdot z$ for any $c > 0$ we ensure that P_G^σ is the Gaussian distribution, because a convolution of two Gaussians is also Gaussian. In particular if we take $G^*(z) = \sqrt{1 - \sigma^2} \cdot z$. Lemma 2 implies that $P_{G^*}^\sigma$ is the standard normal Gaussian $\mathcal{N}(0, 1)$. In

other words, we obtain the global minimum $W_c(P_X, P^\sigma_{G^*}) = 0$ and G^* clearly depends on σ^2. ∎

For the purpose of generative modeling, the noise $\sigma^2 > 0$ is often not desirable, and it is common practice to sample from the trained model G^* by simply returning $G^*(Z)$ for $Z \sim P_Z$ without adding noise to the output. This leads to a mismatch between inference and training. Furthermore, VAE, AVB, and other similar variational methods implicitly use σ^2 as a factor to balance the ℓ_2 reconstruction cost and the KL-regularizer.

In contrast, Proposition 1 shows that for *the same Gaussian models* with *any* given $\sigma^2 \geq 0$ we can minimize $W_c^\dagger(P_X, P^\sigma_G)$ and the solution $G^\dagger$ will be indeed the one resulting in the smallest 2-Wasserstein distance between P_X and the noiseless implicit model $G(Z)$, $Z \sim P_Z$ used in practice.

Blurriness of VAE and AVB We next add to the discussion regarding the blurriness commonly attributed to VAE samples. Our argument shows that VAE, AVB, and other methods based on the marginal log-likelihood *necessarily* lead to an averaging in the input space if $P_G(Y|Z)$ are Gaussian.

First we notice that in the VAE and AVB objectives, for any fixed encoder $Q(Z|X)$, the decoder is minimizing the expected ℓ_2-reconstruction cost $\mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\left[\|X - G(Z)\|^2\right]$ with respect to G. The optimal solution G^* is of the form $G^*(z) = \mathbb{E}_{P^*_z}[X]$, where $P^*_z(X) \propto P_X(X)Q(Z = z|X)$. Hence, as soon as $\mathbf{supp}\, P^*_z$ is non-singleton, the optimal decoder G^* will end up averaging points in the input space. In particular this will happen whenever there are two points x_1, x_2 in $\mathbf{supp}\, P_X$ such that $\mathbf{supp}\, Q(Z|X = x_1)$ and $\mathbf{supp}\, Q(Z|X = x_2)$ overlap.

This overlap necessarily happens in VAEs, which use Gaussian encoders $Q(Z|X)$ supported on the entire $\mathcal{Z}$. When probabilistic encoders Q are allowed to be flexible enough, as in AVB, for any fixed $P_G(Y|Z)$ the optimal Q^* will *try to invert the decoder* (see Appendix D) and take the form

$$Q^*(Z|X) \approx P_G(Z|X) := \frac{P_G(X|Z)P_Z(Z)}{P_G(X)}. \tag{12}$$

This approximation becomes exact in the nonparametric limit of Q. When $P_G(Y|Z)$ is Gaussian we have $p_G(y|z) > 0$ for all $y \in \mathcal{X}$ and $z \in \mathcal{Z}$, showing that $\mathbf{supp}\, Q^*(Z|X = x) = \mathbf{supp}\, P_Z$ for all $x \in \mathcal{X}$. This will again lead to the overlap of encoders if $\mathbf{supp}\, P_Z = \mathcal{Z}$. In contrast, the optimal encoders of WAE do not necessarily overlap, as they are not inverting the decoders.

The common belief today is that the blurriness of VAEs is caused by the ℓ_2 reconstruction cost, or equivalently by the Gaussian form of decoders $P_G(Y|Z)$. We argue that it is instead caused by the *combination* of (a) Gaussian decoders and (b) the objective (KL-divergence) being minimized.

C.2 The 1-Wasserstein distance: relation to WGAN

We have shown that the D_{WAE} criterion leads to a generalized version of the AAE algorithm and can be seen as a relaxation of the optimal transport cost W_c. In particular, if we choose c to be the Euclidean distance $c(x, y) = \|x - y\|$, we get a primal formulation of W_1. This is the same criterion that WGAN aims to minimize in the dual formulation (see Eq. 2). As a result of Theorem 1, we have

$$W_1(P_X, P_G) = \inf_{Q: Q_Z = P_Z} \mathbb{E}_{X\sim P_X, Z\sim Q(Z|X)}\left[\|X - G(Z)\|\right] = \sup_{f\in\mathcal{F}_L} \mathbb{E}_{P_X}\left[f(X)\right] - \mathbb{E}_{P_Z}\left[f(G(Z))\right].$$

This means we can now approach the problem of optimizing W_1 in two distinct ways, taking gradient steps either in the primal or in the dual forms. Denote by Q^* the optimal encoder in the primal and f^* the optimal witness function in the dual. By the envelope theorem, gradients of W_1 with respect to G can be computed by taking a gradient of the criteria evaluated at the optimal points Q^* or f^*.

Despite the theoretical equivalence of both approaches, practical considerations lead to different behaviours and to potentially poor approximations of the real gradients. For example, in the dual formulation, one usually restricts the witness functions to be smooth, while in the primal formulation, the constraint on Q is only approximately enforced. We will study the effect of these approximations.

Imperfect gradients in the dual (i.e., for WGAN) We show that (i) if the true optimum f^* is not reached exactly (no matter how close), the effect on the gradient in the dual formulation can

be arbitrarily large, and (ii) this also holds when the optimization is performed only in a restricted class of smooth functions. We write the criterion to be optimized as $J_D(f) := \mathbb{E}_{P_X}[f(X)] - \mathbb{E}_{P_Z}[f(G(Z))]$ and denote its gradient with respect to G by $\nabla J_D(f)$. Let $\mathcal{H}$ be a subset of the 1-Lipschitz functions $\mathcal{F}_L$ on $\mathcal{X}$ containing smooth functions with bounded Hessian. Denote by $f^*_{\mathcal{H}}$ the minimizer of J_D in $\mathcal{H}$. $A(f, f') := cos\left(\nabla J_D(f), \nabla J_D(f')\right)$ will denote the cosine of the angle between the gradients of the criterion at different functions.

Proposition 2 *There exists a constant $C > 0$ such that for any $\epsilon > 0$, one can construct distributions P_X, P_G and pick witness functions $f_\epsilon \in \mathcal{F}_L$ and $h_\epsilon \in \mathcal{H}$ that are ϵ-optimal $|J_D(f_\epsilon) - J_D(f^*)| \leq \epsilon$, $|J_D(h_\epsilon) - J_D(h^*)| \leq \epsilon$, but which give (at some point $z \in \mathcal{Z}$) gradients whose direction is at least C-wrong: $A(f_\epsilon, f^*) \leq 1 - C$, $A(h_0, h^*) \leq 1 - C$, and $A(h_\epsilon, h^*) \leq 0$.*

Proof We give a sketch of the proof. Consider discrete distributions P_X supported on two points $\{x_0, x_1\}$, and P_Z supported on $\{0, 1\}$ and let $y_0 = G(0)$, $y_1 = G(1)$ ($y_0 \neq y_1$). Given an optimal f^*, one can modify locally it around y_0 without changing its Lipschitz constant such that the obtained f_ϵ is an ϵ-approximation of f^* whose gradients at y_0 and y_1 point in directions arbitrarily different from those of f^*. For smooth functions, by moving y_0 and y_1 away from the segment $[x_0, x_1]$ but close to each other $\|y_0 - y_1\| \leq K\epsilon$, the gradients of f^* will point in directions roughly opposite but the constraint on the Hessian will force the gradients of $f_{\mathcal{F},0}$ at y_0 and y_1 to be very close. Finally, putting y_0, y_1 on the segment $[x_0, x_1]$, one can get an $f^*_{\mathcal{F}}$ whose gradients at y_0 and y_1 are exactly opposite, while taking $f_{\mathcal{F},\epsilon}(y) = f^*_{\mathcal{F}}(y + \epsilon)$, we can swap the direction at one of the points while changing the criterion by less than ϵ. ∎

Imperfect posterior in the primal (i.e., for WAE) In the primal formulation, when the constraint is violated, that is the aggregated posterior Q_Z is not matching P_Z, there can be two kinds of negative effects: (i) the gradient of the criterion is only computed on a (possibly small) subset of the latent space reached by Q_Z; (ii) several input points could be mapped by $Q(Z|X)$ to the same latent code z, thus giving gradients that encourage $G(z)$ to be the average/median of several inputs (hence encouraging a blurriness).

D Further details on VAEs and GANs

VAE, KL-divergence and a marginal log-likelihood For models P_G of the form (3) and *any* conditional distribution $Q(Z|X)$ it can be easily verified that

$$\begin{aligned}-\mathbb{E}_{P_X}[\log P_G(X)] = &- \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_G(Z|X)\big)\right] \\ &+ \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_Z\big) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right]. \end{aligned} \tag{13}$$

Here the conditional distribution $P_G(Z|X)$ is induced by a joint distribution $P_{G,Z}(X, Z)$, which is in turn specified by the 2-step latent variable procedure: (a) sample Z from P_Z, (b) sample X from $P_G(X|Z)$. Note that the first term on the r.h.s. of (13) is always non-positive, while the l.h.s. does not depend on Q. This shows that if conditional distributions Q are not restricted then

$$-\mathbb{E}_{P_X}[\log P_G(X)] = \inf_Q \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_Z\big) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right],$$

where the infimum is achieved for $Q(Z|X) = P_G(Z|X)$. However, for any restricted class $\mathcal{Q}$ of conditional distributions $Q(Z|X)$ we only have

$$\begin{aligned}&- \mathbb{E}_{P_X}[\log P_G(X)] \\ &= \inf_Q -\mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_G(Z|X)\big)\right] + \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_Z\big) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] \\ &\leq \inf_{Q \in \mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_Z\big) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] = D_{\mathrm{VAE}}(P_X, P_G),\end{aligned}$$

where the inequality accounts for the fact that $Q(Z|X)$ might be not flexible enough to match $P(Z|X)$ for all values of X.

Relation between AAE, AVB, and VAE

Proposition 3 *For any distributions P_X and P_G:*

$$D_{\mathrm{AAE}}(P_X, P_G) \leq D_{\mathrm{AVB}}(P_X, P_G).$$

Proof By Jensen's inequality and the joint convexity of D_{GAN} we have

$$\begin{aligned} D_{\mathrm{AAE}}(P_X, P_G) &= \inf_{Q(Z|X)\in\mathcal{Q}} D_{\mathrm{GAN}}\left(\int_{\mathcal{X}} Q(Z|x)p_X(x)dx, P_Z\right) - \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)] \\ &\leq \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{GAN}}\left(Q(Z|X)p_X(x)dx, P_Z\right) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] \\ &= D_{\mathrm{AVB}}(P_X, P_G). \end{aligned}$$

∎

Under certain assumptions it is also possible to link D_{AAE} to D_{VAE}:

Proposition 4 *Assume $D_{\mathrm{KL}}(Q(Z|X), P_Z) \geq 1/4$ for all $Q \in \mathcal{Q}$ with P_X-probability 1. Then*

$$D_{\mathrm{AAE}}(P_X, P_G) \leq D_{\mathrm{VAE}}(P_X, P_G).$$

Proof We already mentioned that $D_{\mathrm{GAN}}(P, Q) \leq 2 \cdot D_{\mathrm{JS}}(P, Q) - \log(4)$ for any distributions P and Q. Furthermore, $D_{\mathrm{JS}}(P, Q) \leq \frac{1}{2} D_{\mathrm{TV}}(P, Q)$ (Lin, 1991, Theorem 3) and $D_{\mathrm{TV}}(P, Q) \leq \sqrt{D_{\mathrm{KL}}(P, Q)}$ (Tsybakov, 2008, Eq. 2.20), which leads to

$$D_{\mathrm{JS}}(P, Q) \leq \frac{1}{2}\sqrt{D_{\mathrm{KL}}(P, Q)}.$$

Together with the joint convexity of D_{JS} and Jensen's inequality this implies

$$\begin{aligned} D_{\mathrm{AAE}}(P_X, P_G) &:= \inf_{Q(Z|X)\in\mathcal{Q}} D_{\mathrm{GAN}}\left(\int_{\mathcal{X}} Q(Z|x)p_X(x)dx, P_Z\right) - \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)] \\ &\leq \inf_{Q(Z|X)\in\mathcal{Q}} D_{\mathrm{JS}}\left(\int_{\mathcal{X}} Q(Z|x)p_X(x)dx, P_Z\right) - \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)] \\ &\leq \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{JS}}\left(Q(Z|X), P_Z\right) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] \\ &\leq \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\left[\frac{1}{2}\sqrt{D_{\mathrm{KL}}\left(Q(Z|X), P_Z\right)} - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] \\ &\leq \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\left(Q(Z|X), P_Z\right) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] \\ &= D_{\mathrm{VAE}}(P_X, P_G). \end{aligned}$$

∎

E FURTHER DETAILS ON EXPERIMENTS

MNIST: We use mini-batches of size 100, $\sigma_z^2 = 1$, and 4x4 convolutional filters. The reported models were trained for 100 epochs. We used $\alpha = 10^{-3}$ for Adam in the beginning, decreased it to 5×10^{-4} after 30 epochs, and to 10^{-4} after first 50 epochs.

CelebA: We pre-processed CelebA images by first taking a 140x140 center crops and then resizing to the 64x64 resolution. We used mini-batches of size 128 and trained the models for various number of epochs (up to 250). All reported models were trained for 60 epochs. Initial learning rate of Adam was set to $\alpha = 10^{-4}$ as often recommended in the literature, decaying by a factor of 1.4 every time the loss plateaus starting from 30th epoch.